

Automated Scoring of Student Essays for Diagnostic Assessments

Statistics & Probability for Data Analytics

Jason M. Bryer, Angela L. Lui, Anthony Fraser, & Julia Ferris

April 24, 2025

Introduction

- In the last few years there have been significant advances in the analysis of text data, specifically the introduction of Large Language Models (LLMs, e.g. OpenAI, Anthropic, etc.).
- The development of LLMs has been largely driven by *transformer* models (Vaswani et al, 2017).
- The Diagnostic Assessment and Achievement of College Skills has been using LightSide (Mayfield, Adamson & Rosé, 2015) to score essays.

Research Question

Do the more modern tokenizers perform better than the traditional rule-based tokenizers?

What is DAACS?

- A suite of technological and social support to optimize student learning.
- No-stakes, diagnostic (formative?) assessments in:
 - Self-Regulated Learning
 - Writing
 - Reading Comprehension
 - Mathematics
- Provides students with immediate feedback about strengths and weaknesses along with links to open educational resources (OER).
- Coaches, Academic Advisors, and Instructors can utilize student results to provide more targeted supports.
- Data is used in predictive analytic efforts to increase the accuracy of identifying "at-risk" students.



SELF-REGULATED LEARNING



COMPLETED



[View previous results](#)



WRITING



COMPLETED



[View previous results](#)



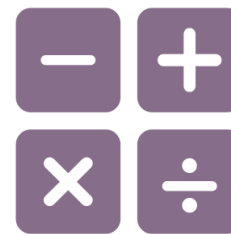
READING



COMPLETED



[View previous results](#)



MATHEMATICS



COMPLETED



[View previous results](#)





Welcome to the Self-Regulated Learning (SRL) assessment! The purposes of this assessment are to:

- generate a profile of your SRL skills - motivation beliefs, self-efficacy, strategic behaviors, and metacognitive skills
- give you feedback about your SRL skills and suggestions for improving them, as needed;
- direct you to online resources and strategies to further enhance your ability to manage the demands of college life.

You will be asked to respond to a series of statements about your SRL skills. For some items, you will indicate your level of agreement, whereas for others you will rate how frequently you display certain behaviors. After this SRL assessment, you will have the opportunity to review your SRL skills and the recommended strategies for improving them, as needed

DAACS Self-Regulated Learning Results

RETAKES ASSESSMENT

Motivation	
Strategies	
Metacognition	
Self-Efficacy (Confidence)	

myDAACS » Self-Regulated Learning

Overview

Step 1

GENERATE DAACS SUMMARY REPORT

Step 2

DOWNLOAD REPORT

Thank you for completing the Self-Regulated Learning (SRL) component of the Diagnostic Assessment and Achievement of Colleges (DAACS). The results presented here are meant to be informative and instructional; **there is no passing or failing**. Because your results can change due to life circumstances or courses you take, you are encouraged to take SRL assessment again, especially when beginning new courses.

[Click here to complete a two question survey to help us make DAACS Self-Regulated Learning section better.](#)



DAACS Writing Prompt



Assessment: **Writing**

Menu ▾

Writing

You received information about your learning skills after you took the self-regulated learning (SRL) survey, as well as suggestions for becoming a more effective and efficient learner. Now, in order to reflect on your learning skills and receive feedback on your writing, please use the results from your SRL survey to do your best writing in a brief essay that answers the questions below.

You will need to refer to your SRL survey results and feedback in your essay. We recommend reviewing them, taking notes, and then returning here to write.

Click here to open your SRL results in a new window.

Essays must be at least 350 words in order to be meaningfully scored. Please aim to write a complete, well-developed essay in order to get accurate feedback about how ready you are for academic writing, and what you can do to strengthen your writing skills.

- What do your self-regulated learning survey results and the feedback tell you about your learning skills? Use results from the survey and the feedback to support your analysis.
- Which suggested strategies from the feedback are you committed to using this term? Explain why you are committed to using those strategies.

Click on Help, then Rubric to review the criteria.

Once you complete the writing assessment, you will be provided feedback and suggestions for resources to use to boost your writing skills, as needed. It is really important that you look at the feedback, because research shows that students who read it and use the resources are more successful than those who simply complete the assessments without viewing the feedback.

Helpful Tips

- What do your self-regulated learning survey results and the feedback tell you about your learning skills? Use results from the survey and the feedback to support your analysis.
- Which suggested strategies from the feedback are you committed to using this term? Explain why you are committed to using those strategies.
- The assessment is not timed and you can stop and return later to complete it.
- You can view the directions and scoring criteria by clicking the *help* button while writing the essay.

DAACS Writing Results



WRITING

Show Results From

September 25, 2024 at 11:08 AM ▾

RETAKE ASSESSMENT

Content	
Organization	
Paragraphs	
Sentences	
Conventions	

Congratulations, you have completed the Writing assessment!

myDAACS » Writing

Overview

Step 1

GENERATE DAACS SUMMARY REPORT

Step 2

DOWNLOAD REPORT

The writing assessment serves three important purposes. It provides an opportunity to reflect on your self-regulated learning survey results, gives you feedback about your writing and suggestions for improving it (as needed), and directs you to online resources that can help you strengthen your writing skills. Your results are below. Click through each section for information about the rubric used to score your writing, as well as suggestions for improvement and links to useful resources. Although we have taken steps to ensure that the scoring of DAACS essays is accurate, some essays are harder to evaluate than others. If you question the accuracy of your results, feel free to reach out to your student mentor, advisor, or the writing center to discuss your writing.



DAACS Writing Feedback Rubric

Sentences	
Grammatically Correct	
Complexity	
Conventions	
Usage	
Punctuation	
← Assessment Overview	

Summary

[MORE INFO](#)

A summary typically contains the main ideas or information from a source in the writer's own words (Yagelski, 2015). In the case of the DAACS writing assessment, a good summary uses relevant survey results and feedback to provide a clear sense of both your strengths and weaknesses in terms of SRL. This summary should lay a foundation for the suggestions for improvement. A summary should not be confused with a paraphrase. The main difference between a summary and a paraphrase is that a summary boils a source text down into a shorter version, whereas a paraphrase restates the source text without necessarily condensing it. Typically, a summary conveys only the main ideas and is shorter than a paraphrase.

	Mastering	Emerging	Developing
Summary	The essay uses relevant survey results and feedback to provide a detailed summary of the student's strengths and weaknesses in terms of self-regulated learning.	<p>The essay refers to survey results and feedback in the summary of strengths and weaknesses related to self-regulated learning.</p> <p>The summary is a bit thin though: Additional, relevant details and examples would have strengthened it.</p>	The summary of strengths and weaknesses in SRL is under-developed. It might need more detail about the results and feedback, and/or should address both strengths and weaknesses, rather than one or the other.

Your score indicates that your essay used relevant survey results and feedback to provide a detailed summary of your strengths and weaknesses in terms of SRL. If you are interested in making your summaries even better, we recommend these resources: <https://owl.excelsior.edu/research/drafting-and-integrating/drafting-and-integrating-summarizing/>

Suggestions

[MORE INFO](#)

Rubric (cont.)

Criteria	SubCriteria	Mastering	Emerging	Developing
content	summary	The essay uses relevant survey results and feedback to provide a detailed summary of the student's strengths and weaknesses in terms of self-regulated learning.	The essay refers to survey results and feedback in the summary of strengths and weaknesses related to self-regulated learning. The summary is a bit thin though: Additional, relevant details and examples would have strengthened it.	The summary of strengths and weaknesses in SRL is under-developed. It might need more detail about the results and feedback, and/or should address both strengths and weaknesses, rather than one or the other.
content	suggestions	The discussion of suggestions for improvement in SRL are logically and explicitly related to the survey results and feedback, and developed in sufficient depth.	The essay includes a clear discussion of improvements to be made to SRL, but the connections to the survey results and feedback are not always explicit and clear.	The discussion of possible improvements in SRL is vague and/or disconnected from the strengths and weaknesses discussed in the summary. It might have been stronger if the suggestions for improvement were more specific, and explicitly tied to the results and feedback.
organization	structure	The essay is well-organized, with an order and structure that present the discussion in a clear, logical manner.	The essay has a basic structure but does not offer a clear, overall organization that enables a reader to understand the progression of one idea to another. It might have been strengthened by the inclusion of an introduction, topic sentences, conclusion, etc.	The essay is not carefully structured. It might move almost randomly from one point to the next. It could have been strengthened by the use of an introduction, topic sentences, a conclusion, and other organizational features.

Rubric (cont.)

Criteria	SubCriteria	Mastering	Emerging	Developing
organization	transitions	Transitions between paragraphs are appropriate and effective, and strengthen the progression of the essay.	Transitions between paragraphs include language that signals a shift from one main idea to the next, as needed (e.g. "In terms of monitoring what I do..." "In the mindset category . . ."). The use of additional transitions (e.g. similarly, in addition, next, as a result) would have helped link the paragraphs and strengthen the overall organization of the essay.	Transitions between paragraphs are missing or ineffective. As a result, the paragraphs abruptly shift from one idea to the next. Transitions (e.g. similarly, in addition, next, as a result) would have helped link the paragraphs and strengthen the overall organization of the essay.
paragraphs	ideas	Paragraphs are consistently and clearly focused on a main idea or point.	In general, each paragraph has a clear focus. Some might have been strengthened by a topic sentence that establishes the focus, but the main ideas are apparent.	Most or all paragraphs lack a clear focus on a main point or topic, perhaps because too many ideas are packed into one paragraph. Each topic should be fully developed in separate paragraphs. Each paragraph should include a statement of the main point, and every sentence in the paragraph should relate to that point.
paragraphs	cohesion	Within paragraphs, the individual sentences are seamlessly linked together; the reader can see the relationship between the ideas or information in one sentence and those in another sentence. The writing explicitly links sentences and ideas using adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which), conjunctions (e.g., and, or, while, whereas), and/or	The ideas or information in each sentence within a paragraph are loosely linked together. The connections between sentences would have been clarified by additional or better choices of linking words and phrases, such as adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which), conjunctions (e.g., and, or, while, whereas), and/or the repetition of key	Because the ideas or information in each sentence within a paragraph are not linked together, it is hard for a reader to see the relationship between sentences. Ideas could be connected with adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which), conjunctions (e.g., and, or, while, whereas), and/or the repetition of key words.

Rubric (cont.)

Criteria	SubCriteria	Mastering	Emerging	Developing
sentences	correct	Sentences are correct: no run-ons, fragments, or errors in subject-verb agreement.	The sentences are generally correct. Minor issues with grammar, such as errors in subject-verb agreement (e.g., "the survey results suggests...") do not interfere with meaning.	Problems with grammar make the essay hard to read and understand. More attention to the parts of speech (e.g., nouns, pronouns, verbs, adjectives, etc.) was needed.
sentences	complexity	Consistent and appropriate use of a variety of sentence structures, including some sophisticated, complex sentence structures and syntactic forms.	Complex syntactic structures are present but not consistently used; sentence structure is varied but not often sophisticated.	The sentences lack syntactic complexity and vary little, if at all, in structure. The sentences are generally simple in structure (subject-verb-object).
conventions	conventions	Spelling, punctuation, and capitalization are correct to the extent that almost no editing is needed.	Spelling, punctuation, and capitalization are correct to the extent that only light editing is needed.	Careful editing was needed to correct frequent, distracting errors in spelling, punctuation, and/or capitalization.

Data Source

1,065 essays were collected from Western Governors University and the University at Albany, 592 of which were double scored. For WGU, the first 893 essays that were collected as part of a larger randomized control trial were included. The remaining essays were randomly selected from the University at Albany.

There were two scoring events with a total of 15 raters supervised by two subject matter experts.

Criteria	IRR
Content Summary	0.58
Content Suggestions	0.58
Organization Structure	0.63
Organization Transition	0.58
Paragraphs Focus on a Main Idea	0.62
Paragraphs Cohesion	0.60
Sentences Correct	0.55
Sentences Complex	0.59
Conventions	0.55

Workflow for automated machine scoring

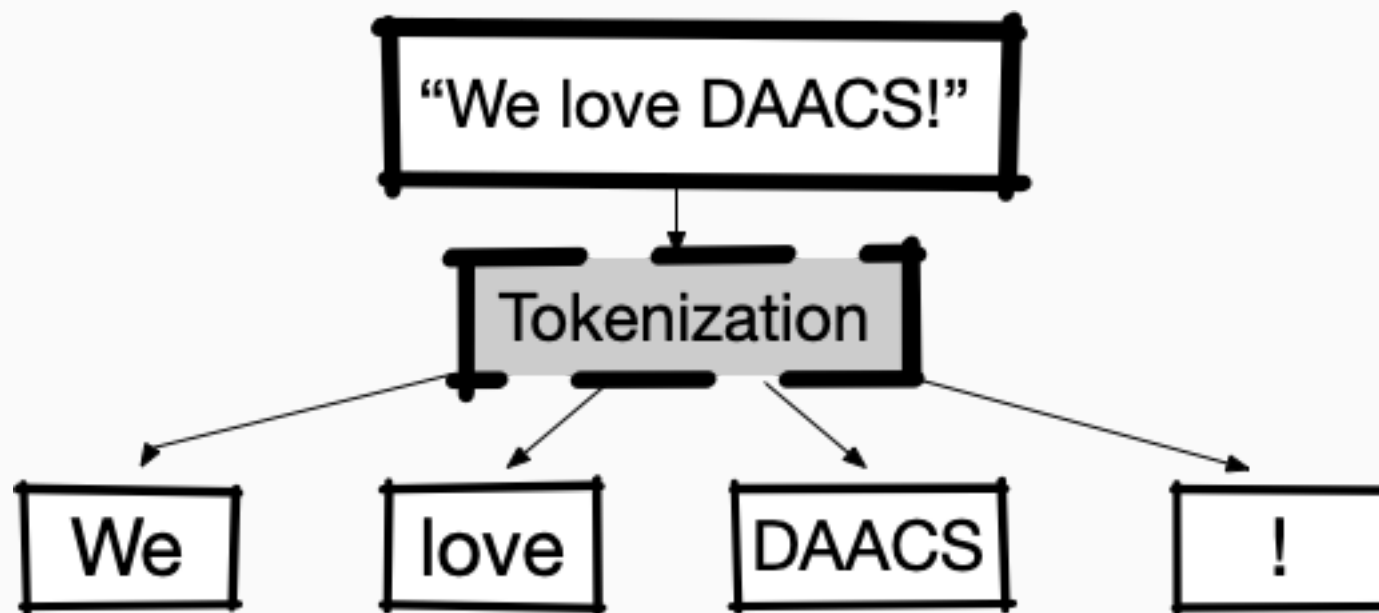


For this study predictive models were trained on 75% of the observations. Prediction accuracies are calculated from the test data.

For essays that were double scored, the conferral score was used. Otherwise the single rater score was used.

Tokenization

Tokenization is the process of converting text into vectors or matrices.



Vaswani et al (2017) introduced the *transformer* model approach to tokenization. Unlike traditional rule based tokenization, transformer models are trained on a large corpus of text in order to not only find the basic word structures, but to also to infer context by finding nearby reoccurring tokens.

Tokenizers and Predictive Models

Tokenization Algorithm	Reference
Bert	Devlin et al. (2018)
Distilbert	Sanh, Debut, Chaumond, and Wolf (2019)
Word 2 Vec	Mikolov, Chen, Corrado, and Dean (2013)
Scikit Native Vectors	Arsenovic et al. (2022)
Facebook Vectors	Bojanowski et al. (2016)
Scikit Vectors	Arsenovic et al. (2022)

Prediction Model	Reference
k-nearest neighbors	Fix & Hodges (1951)
Linear Support Vector Classification	Ratna et al (2019)
C-Support Vector Classification	Novakovic et al (2011)
Gradient Boosting	Breiman (1997)
XGBoost	Shang, Men, and Du (2023)
Bagging	Islam et al (2022)
Random Forest	Breimann (2001)
AdaBoost	Sevinç (2022)
Logistic Regression	Cramer (2002)

Version 1: LightSide Features

Feature	Summary	Suggestions	Structure	Transition	Focus	Cohesion	Correct	Complex	Conventions
Model	Logit	Logit	Bayes	Bayes	Logit	Logit	Logit	Bayes	Logit
Unigrams	X	X	X	X	X	X	X	X	X
Bigrams					X		X	X	
Trigrams							X	X	
POS Bigrams							X	X	X
POS Trigrams							X	X	X
Word/POS Pairs	X	X	X	X	X	X	X	X	X
Line Length		X	X	X	X	X	X	X	X
Include Punctuation	X	X	X	X	X	X	X	X	X
Step N-Grams				X					
Character N-Grams			X	X	X	X		X	
Stretchy Patterns			X	X		X	X	X	X

LightSide (Mayfield, Adamson & Rosé, 2015)

Version 1: IRR and LightSide Accuracy

Criteria	Human-Human	Human-LightSide
Conntent Suggestions	0.6089	0.7226
Content Summary	0.5578	0.6962
Conventions	0.5533	0.6316
Organization Structure	0.6289	0.7422
Organization Transitions	0.5711	0.4717
Paragraphs Cohesion	0.6156	0.7273
Paragraphs Focus on a Main Ideas	0.5978	0.7345
Sentences Complexity	0.5600	0.6842
Sentences Correct	0.5633	0.5573

Results



Results: Top performing models by criteria

Criteria	Toeknizer	Predictive Model	Human-LLM	Human-Human	Human-LightSide	Improvement
Content Suggestions	scikit_vectors	LinearSVC	0.71	0.61	0.72	-0.01
Content Summary	bert_vectors	RandomForest	0.72	0.56	0.70	0.02
Conventions	bert_vectors	GradientBoosting	0.71	0.55	0.63	0.08
Organization Structure	bert_vectors	RandomForest	0.77	0.63	0.74	0.03
Organization Transition	bert_vectors	RandomForest	0.66	0.57	0.47	0.19
Paragraphs Cohesion	distilbert_vectors	RandomForest	0.75	0.62	0.73	0.03
Paragraphs Focus on a Main Idea	distilbert_vectors	RandomForest	0.73	0.60	0.73	-0.01
Sentences Complex	distilbert_vectors	GradientBoosting	0.73	0.56	0.68	0.04
Sentences Correct	bert_vectors	RandomForest	0.70	0.56	0.56	0.15

Discussion

- In general, the new generation of tokenizers are an improvement over the rule based tokenizers.
- The more modern tokenizers (e.g. BERT, SciKit Vectors, word2vec, etc.) all perform equally well.
- The choice of predictive model seems to have a bigger impact on model accuracy. This is consistent with prior research (see e.g. [Fernández-Delgado, Cernados, Barro, & Amorin, 2014](#)).
- These new models will soon be integrated into the DAACS system.

Thank You!

✉ jason.bryer@cuny.edu

🐙 @DAACS

🔗 www.daacs.net



DAACS was developed under grants #P116F150077 and #R305A210269 from the U.S. Department of Education. However, the contents do not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal Government.